

Decoding Market Emotions in Cryptocurrency Tweets via Predictive-Statement Classification with Machine Learning and Transformers

Moein Shahiki Tash¹, Zahra Ahani¹, Mohim Tash², Mostafa Keikhay Farzaneh²,
Ari Y. Barrera-Animas³, Olga Kolesnikova^{1,*}

¹ Instituto Politécnico Nacional,
Centro de Investigación en Computación, Mexico City,
Mexico

² University of Sistan and Baluchestan,
Faculty of Management and Economics, Zahedan,
Iran

³ Universidad Panamericana, Mexico City,
Mexico

mshahikit2022@cic.ipn.mx, kolesnikova@cic.ipn.mx

Abstract. Cryptocurrency discussions on social media combine emotional reactions with statements about future market behavior, but these two signals are not equivalent. This study presents a two-stage framework for identifying predictive statements in English-language cryptocurrency posts concerning Cardano (ADA), Binance Coin (BNB), Polygon (MATIC), XRP, and Fantom (FTM). Task 1 distinguishes Predictive from Non-Predictive posts; Task 2 assigns Predictive posts to Incremental, Decremental, or Neutral categories. The original corpus contains 3,116 human-reviewed posts. Two fixed, stratified 80/20 splits were created independently before augmentation, one for the binary Task 1 labels and one for the three-way Task 2 direction labels, leaving 624 original posts for Task 1 testing and 224 original predictive posts for Task 2 testing. GPT-generated paraphrases were added only to minority classes in the training partition. We compared TF-IDF-based traditional classifiers, CNN and BiLSTM models with GloVe or FastText embeddings, and five transformer checkpoints. Macro-F1 was the primary metric. Under the augmented-training condition, the XLM-RoBERTa language-identification checkpoint achieved the highest Task 1 macro-F1 (0.7011), whereas Random Forest achieved the highest Task 2 macro-F1 (0.6488), narrowly exceeding SVM-RBF (0.6478). SenticNet was used only after classification

to describe non-exclusive emotion distributions; it was not used as a classifier input. Positive and negative emotions appeared across directional classes, showing why emotional polarity alone cannot reliably represent market expectations. The results support task-dependent model selection and a strict separation between directional prediction and post-hoc emotion analysis.

Keywords. Cryptocurrency, predictive statements, social media, text classification, data augmentation, emotion analysis, SenticNet.

1 Introduction

Cryptocurrencies are digital assets whose transactions are recorded on distributed ledgers and whose markets are characterized by rapid information diffusion, speculative participation, and pronounced volatility [5, 9, 15]. Bitcoin, introduced by Nakamoto in 2008, relies on proof-of-work consensus and a public transaction ledger [7]. Its adoption helped establish a broader ecosystem of digital assets and attracted investors, regulators, fund managers, and researchers.

Social media has become an important venue for cryptocurrency discussion. Users report events, react to price movements, express emotions, and make explicit or implicit forecasts. These functions are often conflated. A positive post does not necessarily predict a price increase, and a negative post does not necessarily predict a decline. For example, a user may express enthusiasm about a predicted price decrease because it creates a perceived buying opportunity, while another user may express frustration during an upward market because an opportunity was missed. Treating affective polarity as a direct proxy for expected direction can therefore produce misleading signals [3, 14].

This study addresses that distinction through predictive-statement classification. We define four mutually exclusive labels: Incremental predictions, which anticipate an upward movement or improvement; Decremental predictions, which anticipate a decline; Neutral predictions, which anticipate stability or no material change; and Non-Predictive statements, which do not assert an expected future outcome. The classification pipeline is hierarchical. Task 1 detects predictive intent, and Task 2 determines the direction of posts classified as predictive.

The study makes four contributions. First, it operationalizes predictive intent separately from emotional tone. Second, it evaluates traditional machine-learning, neural, and transformer models under the same two-stage task formulation. Third, it addresses class imbalance with a split-before-augmentation protocol in which synthetic paraphrases are confined to the training partition. Fourth, it uses SenticNet only for post-hoc interpretation, enabling emotion distributions to be examined without creating a circular relationship between emotion features and directional labels.

2 Task Definition

2.1 Predictive and Non-predictive Statements

A predictive statement asserts or clearly implies an expected future outcome for a cryptocurrency or the broader market. A non-predictive statement reports a past or present event, expresses a

general opinion without a future claim, or asks a speculative question without indicating the author's expected answer.

2.2 Directional Classes

An **Incremental** prediction anticipates growth, improvement, an upward price movement, or the crossing of a higher threshold. A **Decremental** prediction anticipates decline, deterioration, a downward price movement, or the crossing of a lower threshold. A **Neutral** prediction anticipates stability, consolidation, sideways movement, or no substantial change.

Table 1 provides representative examples. The Neutral examples explicitly predict stability; statements that merely express uncertainty are not treated as Neutral predictions.

Table 1. Representative examples of the annotation categories.

Category	Example
Incremental	ADA will probably break \$1.20 next week if this momentum continues.
Incremental	Profits are expected to rise during the next quarter.
Decremental	I think BTC is going to drop below \$20,000 soon.
Decremental	Sales are projected to decline in the next fiscal quarter.
Neutral	MATIC looks stable; I do not expect a major movement this week.
Neutral	The price is expected to remain close to its current level next month.
Non-Predictive	XRP announced a new partnership today.
Non-Predictive	Blockchain technology is used in several industries.

3 Related Work

Cryptocurrency forecasting research has largely focused on numerical market series, technical indicators, sentiment signals, or combinations of these sources. Vieitez et al. [16] used contextual market and online features with recurrent neural networks for Ethereum price regression and support-vector machines for trend classification. Qureshi et al. [10] compared classifiers for algorithmic trading with technical indicators, imbalance correction, and backtesting. Kumar et al. [6] combined LSTM and SVM components for stock and cryptocurrency prediction, while Golnari et al. [4] proposed probabilistic recurrent models and transfer learning to quantify uncertainty across multiple assets. Otabek and Choi [8] reviewed the relationship between forecasting accuracy and practical trading strategies.

These studies primarily predict market values or movements from numerical or aggregated signals. The present work instead identifies forecasting language in individual social-media posts. This linguistic layer can complement time-series systems by distinguishing an asserted expectation from a contemporaneous emotional reaction.

Prior work on the same cryptocurrency domain used SenticNet to examine emotion trends and their association with prices [14, 13]. SenticNet provides concept-level affective representations that can capture more nuanced information than simple positive-negative polarity [1, 12]. In the present study, however, SenticNet is not part of the predictive classifier. It is applied after classification to characterize how emotion categories co-occur with directional predictions.

4 Materials and Methods

4.1 Data Source

The source corpus was derived from the English-language X/Twitter collection described in [13], covering October 2022 through March 2023. The broader collection contained approximately 30,000 posts. From it, 3,116 posts concerning five cryptocurrency assets were selected for predictive-statement annotation: Cardano (ADA),

Binance Coin (BNB), Polygon (MATIC, the ticker used during the collection period), XRP, and Fantom (FTM).

An automated pass targeted exact duplicates and near-duplicate reposts before annotation; a post-hoc audit of the released corpus nonetheless found that verbatim-duplicate post text remained in a portion of the retained data, in some cases carrying conflicting labels (Section 8). User identifiers and other personally identifying fields were excluded from the modeling dataset.

4.2 Annotation Protocol

Two English-proficient annotators with master's-level training in computer science completed a calibration exercise on 100 examples before labeling the corpus. The decision process was hierarchical:

1. Determine whether the post asserts or clearly implies a future outcome.
2. If it is predictive, assign Incremental, Decremental, or Neutral according to the expected direction.
3. If several forecasts occur in one post, assign the dominant or most strongly emphasized forecast.
4. Do not label a purely speculative question as predictive unless the author also signals an expected outcome.
5. Treat sarcasm or ambiguity conservatively; assign a predictive label only when the future claim and direction are recoverable from context.

Every original post retained in the dataset had a human-reviewed final label. GPT-4o was used as an auxiliary annotator and as a consistency check, not as an automatic replacement for human judgment.

Agreement between GPT-4o outputs and human-reviewed labels for Task 1 was Cohen's $\kappa = 0.7493$, indicating substantial agreement. This coefficient describes human-AI agreement for binary predictive-statement detection; it is not presented as a human-human reliability coefficient or as evidence for Task 2 agreement.

4.3 Model-specific Preprocessing

Preprocessing was separated by model family. For TF-IDF and static-embedding models, URLs and user mentions were normalized, repeated whitespace was collapsed, and text was tokenized. Short but semantically important tokens, including negators and directional terms such as *no*, *up*, and *go*, were retained. For transformer models, the text was kept close to its natural form: URLs and handles were normalized, while punctuation and word order were preserved. Each transformer used its native tokenizer. This separation avoids applying aggressive traditional preprocessing to pretrained contextual models.

4.4 Fixed Train-test Split and Augmentation

Two fixed, stratified 80/20 splits were used, each created with a fixed random seed before any synthetic text was generated. The Task 1 split stratifies the complete 3,116-post corpus on the binary Predictive/Non-Predictive label. The Task 2 split independently stratifies the 1,116-post predictive subset on the three-way direction label (Incremental, Decremental, Neutral).

Since these two splits are computed independently rather than derived from a single joint assignment, the predictive-class counts differ by one between them: the Task 1 split places 893 predictive posts in training and 223 in the held-out test set, whereas the Task 2 split places 892 in training and 224 in the held-out test set (Table 2). Consequently, an original post held out for Task 1 evaluation is not guaranteed to also fall in the Task 2 test partition, and vice versa; the two held-out sets overlap substantially but are not identical. This does not introduce paraphrase leakage in either task, since no synthetic instance was added to either test partition, and every model within a given task was fit and evaluated on that task's own fixed partition only. Model fitting and any configuration decisions used only the relevant training partition, and each held-out test partition was evaluated once after training.

Table 2 reports the exact allocation. Within each task, the integer class counts can differ by one from a simple per-class rounding of 20% because the split preserves class proportions

across the complete task-specific dataset rather than rounding each class independently.

GPT-based paraphrasing was used to balance only the training partition. Synthetic instances were generated for minority classes until each class matched the largest original training class within its task. The prompt instructed the model to preserve the original predictive direction, financial context, and concise social-media style. Generated texts were linked to their source posts so that source and paraphrase provenance remained auditable. No synthetic instance was added to either test partition.

4.5 Classification Models

Three model families were evaluated.

Traditional machine learning. Logistic Regression, Random Forest, XGBoost, linear-kernel SVM, and RBF-kernel SVM were trained on TF-IDF representations [11]. These models provide strong sparse-text baselines and differ in their assumptions about linearity and feature interactions.

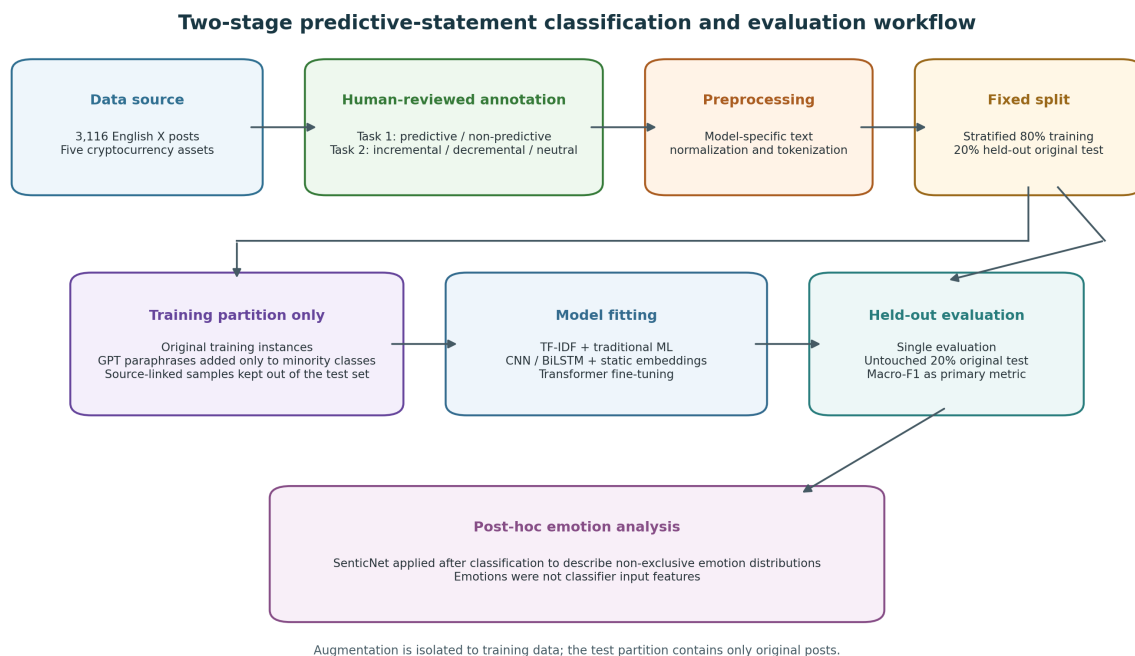
Deep learning. Convolutional neural networks (CNNs) and bidirectional long short-term memory networks (BiLSTMs) were paired with GloVe and FastText word embeddings. CNNs capture local phrase patterns, whereas BiLSTMs model sequential dependencies.

Transformers. The evaluated checkpoints were DistilRoBERTa, RoBERTa-base, BERT-base-uncased, RoBERTuito, and papluca/xlm-roberta-base-language-detection. The last model is reported as **XLM-RoBERTa-LID** to distinguish it from the generic xlm-roberta-base checkpoint. It was included as an exploratory transfer baseline initialized from a language-identification-tuned model. RoBERTuito, which was pretrained primarily on Spanish social-media text, was included as a cross-language, social-media-domain control rather than as an English-specific model. Results from these two checkpoints should therefore be interpreted as transfer results, not as evidence that either is the optimal pretrained model for English cryptocurrency text.

Within each task, the same fixed data partition was used for all model families: every Task 1 model

Table 2. Original, training, held-out test, and augmented-training class counts

Task and label	Original total	Original training (80%)	Held-out test (20%)	Training after augmentation
Task 1: Non-Predictive	2,000	1,599	401	1,599
Task 1: Predictive	1,116	893	223	1,599
Task 1 total	3,116	2,492	624	3,198
Task 2: Incremental	570	456	114	456
Task 2: Decremental	434	347	87	456
Task 2: Neutral	112	89	23	456
Task 2 total	1,116	892	224	1,368

**Fig. 1.** Two-stage predictive-statement classification and evaluation workflow. Augmentation is restricted to training data, and the fixed test partition contains only original posts

was fit and evaluated on the same Task 1 split, and every Task 2 model was fit and evaluated on the same Task 2 split. No test-set observations were used for augmentation, fitting, or model selection.

4.6 Evaluation Metrics

Accuracy, macro-averaged precision, macro-averaged recall, macro-F1, and their support-weighted counterparts were computed from held-out predictions [2]. Macro-F1 was

designated as the primary selection metric because it gives each class equal weight and is less dominated by the majority class.

Accuracy is reported as a secondary measure. For a task with K classes, macro-F1 is the arithmetic mean of the K class-specific F1 scores.

For context, an always-majority classifier on the held-out Task 1 set would obtain accuracy 0.6426 and macro-F1 0.3912. On Task 2 it would obtain accuracy 0.5089 and macro-F1 0.2249.

Table 3. Percentage of posts assigned to each non-exclusive emotion category by cryptocurrency and Task 2 label (Note: Categories are non-exclusive; row percentages are not expected to sum to 100%)

Asset	Task 2 label	Delight & Joy	Enthusiasm & Eagerness	Delight & Pleasantness	Grief & Sadness	Fear & Anxiety	Rage & Anger
BNB	Neutral	36.84	15.79	36.84	5.26	31.58	5.26
BNB	Decremental	44.09	26.88	16.13	24.73	13.98	10.75
BNB	Incremental	47.57	40.78	30.10	12.62	20.39	2.91
MATIC	Decremental	58.33	42.86	45.24	7.14	2.38	2.38
MATIC	Incremental	72.41	49.66	39.31	5.52	1.38	0.69
MATIC	Neutral	68.18	63.64	31.82	4.55	4.55	4.55
ADA	Decremental	59.52	42.86	25.00	13.10	7.14	16.67
ADA	Incremental	61.29	44.09	37.63	15.05	4.30	8.60
ADA	Neutral	46.15	34.62	30.77	19.23	3.85	—
FTM	Decremental	66.67	26.39	20.83	16.67	5.56	6.94
FTM	Incremental	81.82	41.82	26.36	6.36	3.64	2.73
FTM	Neutral	83.33	41.67	37.50	12.50	8.33	4.17
XRP	Incremental	58.33	32.41	32.41	9.26	1.85	6.48
XRP	Decremental	52.69	27.96	32.26	16.13	5.38	11.83
XRP	Neutral	57.14	28.57	52.38	9.52	—	14.29

These baselines illustrate why accuracy alone is insufficient for either task.

4.7 Post-hoc Emotion Analysis

SenticNet [1, 12] was applied only after classification. It was not concatenated with TF-IDF vectors or supplied to any classifier.

For each cryptocurrency-by-Task-2-label group, the percentage for an emotion category was calculated as the number of posts assigned that category divided by the number of posts in the group, multiplied by 100.

The six reported paired categories—Delight and Joy, Enthusiasm and Eagerness, Delight and Pleasantness, Grief and Sadness, Fear and Anxiety, and Rage and Anger—follow the affective grouping used in the source analysis [13].

Categories are non-exclusive: a post may contribute to more than one category, so percentages within a row need not sum to 100%.

The two categories containing *Delight* represent separate predefined pairings and are interpreted independently rather than merged.

5 Results

5.1 Emotion Distributions

Table 3 reports the non-exclusive emotion percentages. Incremental predictions were frequently associated with positive categories, particularly for FTM and MATIC. Positive categories also appeared in Decremental posts; for example, Delight and Joy reached 58.33% for MATIC and 66.67% for FTM in that class. This pattern may be consistent with strategic optimism or “buy-the-dip” discourse, but the emotion labels alone cannot establish that motivation. Confirmation would require contextual or user-level behavioral analysis. Negative categories such as Grief and Sadness and Rage and Anger also occurred in Decremental predictions, illustrating heterogeneous reactions to the same expected direction.

5.2 Traditional Machine Learning Models

Table 4 compares training on the original imbalanced partition with training on its augmented, class-balanced version. On Task 1,

XGBoost obtained the strongest macro-F1 without augmentation (0.5068), while SVM-RBF performed best after augmentation (0.6825).

On Task 2, XGBoost was strongest without augmentation (0.3631). After augmentation, Random Forest achieved the highest macro-F1 (0.6488), narrowly ahead of SVM-RBF (0.6478) and linear SVM (0.6463).

5.3 Deep Learning Models

Among models trained on the original Task 1 partition, BiLSTM with GloVe achieved the highest macro-F1 (0.5070). It remained the strongest neural model after Task 1 augmentation (0.6349).

For Task 2, CNN with FastText led on the original training data (0.3656) and also achieved the highest augmented-training macro-F1 (0.6153), slightly exceeding CNN with GloVe (0.6149).

5.4 Transformer Models

RoBERTa-base produced the highest original-training macro-F1 for Task 1 (0.5578), and DistilRoBERTa was strongest for Task 2 (0.4051). With augmented training, XLM-RoBERTa-LID achieved the best Task 1 macro-F1 (0.7011). DistilRoBERTa was the strongest transformer for Task 2 (0.5936). The strong Task 1 result of XLM-RoBERTa-LID should be interpreted in light of its language-identification fine-tuning, and it should not be conflated with a generic XLM-RoBERTa baseline.

5.5 Cross-family Comparison

The best model depended on the task. XLM-RoBERTa-LID achieved the highest Task 1 macro-F1 (0.7011), followed by SVM-RBF (0.6825). For Task 2, Random Forest achieved the highest macro-F1 (0.6488), closely followed by SVM-RBF (0.6478) and linear SVM (0.6463).

The strongest deep-learning model, CNN with FastText, reached 0.6153, and the strongest transformer, DistilRoBERTa, reached 0.5936. Thus, transformers did not outperform traditional models on both tasks. Sparse lexical and directional cues appear particularly competitive for the three-way distinction.

6 Error Analysis

Because the fixed test partitions were relatively small, error analysis focused on patterns supported by the held-out aggregate results and manual inspection of representative errors.

For Task 1, the principal difficulty involved implicit forecasts that lacked overt future markers such as *will*, *going to*, or *expect*. Conditional statements and posts combining a present observation with an implied future direction were also difficult. Some non-predictive opinions contained bullish or bearish vocabulary but no recoverable future assertion, producing lexical overlap with predictive posts.

For Task 2, Incremental and Decremental predictions often shared the same sentence structure and asset vocabulary, differing only in a small number of polarity or direction cues. Negation, target thresholds, sarcasm, and multi-clause posts created additional ambiguity. The Neutral class was conceptually distinct but small in the original training data, making strict leakage-free augmentation especially important. These observations motivate future analyses with source-grouped augmentation, explicit negation handling, and English financial-social-media pre-trained models.

7 Discussion

The results demonstrate that predictive intent and emotional tone should be modeled as related but separate dimensions. Positive emotions frequently accompanied Incremental predictions, and negative emotions frequently accompanied Decremental predictions, but neither relationship was deterministic. The presence of positive emotions in some Decremental posts is consistent with several possible interpretations, including buying-opportunity discourse, relief, humor, or excitement about short positions. Because SenticNet labels alone cannot distinguish these motivations, the finding should not be treated as direct evidence of “buy-the-dip” behavior.

Model performance was task-dependent. Task 1 required identifying whether a future claim was present, a contextual distinction for which the

Table 4. Traditional model results: macro-F1 with accuracy in parentheses

Model	Task 1 original train	Task 1 augmented train	Task 2 original train	Task 2 augmented train
Logistic Regression	0.4986 (0.6435)	0.6672 (0.6683)	0.3436 (0.5269)	0.6396 (0.6413)
Random Forest	0.4697 (0.5725)	0.6380 (0.6400)	0.3313 (0.5009)	0.6488 (0.6513)
XGBoost	0.5068 (0.5969)	0.6389 (0.6420)	0.3631 (0.5063)	0.6059 (0.6068)
SVM, linear	0.5044 (0.6239)	0.6583 (0.6605)	0.3493 (0.5323)	0.6463 (0.6454)
SVM, RBF	0.4821 (0.6200)	0.6825 (0.6885)	0.3034 (0.5323)	0.6478 (0.6524)

Table 5. Deep-learning results: macro-F1 with accuracy in parentheses

Model	Task 1 original train	Task 1 augmented train	Task 2 original train	Task 2 augmented train
CNN + GloVe	0.4986 (0.5577)	0.6277 (0.6430)	0.3615 (0.5008)	0.6149 (0.6336)
BiLSTM + GloVe	0.5070 (0.5728)	0.6349 (0.6490)	0.3630 (0.5026)	0.5737 (0.5886)
CNN + FastText	0.4816 (0.5430)	0.6231 (0.6425)	0.3656 (0.4954)	0.6153 (0.6254)
BiLSTM + FastText	0.4821 (0.5555)	0.6222 (0.6353)	0.3485 (0.4695)	0.6014 (0.6120)

Table 6. Transformer results: macro-F1 with accuracy in parentheses

Model	Task 1 original train	Task 1 augmented train	Task 2 original train	Task 2 augmented train
DistilRoBERTa	0.5453 (0.5960)	0.6593 (0.6697)	0.4051 (0.5350)	0.5936 (0.6202)
RoBERTa-base	0.5578 (0.6159)	0.6625 (0.6755)	0.3294 (0.5278)	0.5901 (0.6179)
RoBERTa-tiny	0.5345 (0.5818)	0.6444 (0.6575)	0.3817 (0.5090)	0.5505 (0.5758)
BERT-base-uncased	0.5213 (0.5812)	0.6378 (0.6548)	0.3759 (0.4937)	0.5439 (0.5693)
XLM-RoBERTa-LID	0.4301 (0.6486)	0.7011 (0.7127)	0.3396 (0.5260)	0.5367 (0.5640)

strongest transformer performed well. Task 2 required separating directions expressed through overlapping vocabulary and similar structures.

In that setting, TF-IDF-based Random Forest and SVM models were more competitive than the evaluated neural and transformer baselines.

This result does not imply that traditional models are universally superior for fine-grained financial language. Rather, it shows that, for this dataset and these checkpoints, lexical indicators and nonlinear interactions in a sparse feature space were effective.

The augmentation results must also be interpreted within the fixed-split design. Paraphrasing increased minority-class representation in training, while the held-out set remained original and retained its natural imbalance.

This design avoids evaluating models on synthetic text and prevents a source post and its paraphrase from appearing on opposite sides of the train-test boundary. Macro-F1 remains

the central metric because it exposes changes in minority-class performance that accuracy can obscure.

The emotion analysis is descriptive, not causal or predictive of market prices. It characterizes the affective content of posts after directional labels have been determined.

Establishing whether any emotion-prediction combination forecasts subsequent market behavior would require time-aligned prices, preregistered horizons, transaction-cost-aware backtesting, and out-of-period validation.

8 Limitations

Several limitations qualify the findings. First, the dataset contains English posts from a single platform and five cryptocurrency communities, limiting linguistic, temporal, and platform generalizability. Second, Task 2 is relatively small, and the Neutral class contains only 112 original

examples, including 23 in the held-out test set. Class-level estimates for Neutral should therefore be interpreted cautiously.

Third, the reported $\kappa = 0.7493$ measures agreement between GPT-4o and human-reviewed Task 1 labels. A separate human-human agreement coefficient and a Task 2 reliability coefficient were not available in the reported experiment. Future annotation studies should report reliability for each stage, the exact adjudication rate, and confidence intervals.

Fourth, two transformer baselines have important domain qualifications: RoBERTuito is primarily Spanish and XLM-RoBERTa-LID was previously tuned for language identification. Future work should prioritize English Twitter/X and financial-domain checkpoints and evaluate them with repeated independent held-out splits or an external test collection.

Fifth, the Task 1 and Task 2 held-out partitions were produced by two independently computed stratified splits rather than a single joint split of the corpus (Section 4.4), so the predictive-class train/test counts differ by one post between the two tasks and the two held-out sets are not identical. No paraphrased text entered either test partition under this design, but a post excluded from the Task 1 test set is not guaranteed to be excluded from the Task 2 test set, and readers should not interpret the two tasks as sharing one exact partition of the predictive posts. A single joint stratified split across both the binary and directional labels would remove this discrepancy and is recommended for future iterations of this pipeline.

Sixth, a post-hoc audit of the released corpus identified verbatim-duplicate post text that the deduplication pass described in Section 4.1 did not fully remove. Of the 3,116 original posts, 962 (30.9%) share identical text with at least one other post, forming 346 duplicate-text groups. Of these groups, 319 (92%) contain copies of the same text carrying conflicting Task 1 labels (Predictive in one copy and Non-Predictive in another), affecting 904 posts, or 29.0% of the original corpus; a further 13 groups show conflicting Task 2 direction labels among predictive duplicates. Because the train-test split described in Section 4.4 was not constructed with this duplication in mind, a

duplicated and contradictorily labeled post could appear in the training partition under one label and in the held-out test partition under the other, which would bias the reported Task 1 metrics in an unknown direction.

This issue was identified after the experiments reported here were completed; it does not reflect an error in the splitting or augmentation procedures themselves, but it qualifies the reliability of the underlying labels and should be resolved, for example by reconciling each duplicate-text group to a single label or removing conflicting duplicates, before the corpus is reused in future work.

Finally, this study does not validate the identified statements against subsequent market prices and therefore evaluates linguistic classification, not financial forecasting accuracy.

9 Conclusion and Future Work

This study introduced a two-stage framework for detecting predictive statements in cryptocurrency-related social-media text and classifying their direction as Incremental, Decremental, or Neutral.

A fixed stratified split isolated 20% of the original corpus for testing before any augmentation. GPT-generated paraphrases were restricted to the training partition, and SenticNet was reserved for post-hoc emotion analysis.

The strongest Task 1 result was obtained by XLM-RoBERTa-LID (macro-F1 0.7011), while Random Forest produced the strongest Task 2 result (0.6488), closely followed by SVM-RBF (0.6478). CNN with FastText was the strongest deep-learning configuration for Task 2 (0.6153).

These findings reject a one-model-fits-all interpretation and support selecting representations according to the linguistic distinction being modeled.

Future work should expand the annotated corpus, report stage-specific human-human reliability, evaluate English financial-social-media checkpoints, and test temporal transfer to later market periods. A separate forecasting study could then align predictive statements and post-hoc emotion profiles with subsequent price movements under a prospectively defined evaluation protocol.

Declarations

Data Availability

The study uses derived labels from public social-media posts. To respect platform terms and user privacy, redistribution should be limited to post identifiers and derived annotations where permitted. The annotation schema, prompts, and analysis code can be made available with the article or through the corresponding author, subject to the platform's terms of service.

Declaration of Generative AI Use

GPT-4o was used as an auxiliary annotator and GPT-based paraphrasing was used to augment minority classes in the training partition. AI-generated outputs were subject to human review and were not used in the held-out test set. The authors remain responsible for the study design, labels, analysis, and conclusions.

Acknowledgments

The authors thank the participating annotators and the supporting institutions named in the Funding statement. The work was done with partial support from grant 20260496 (O.K.) by Secretary of Research and Posgraduate Studies (SIP) of Instituto Politécnico Nacional, Mexico.

References

1. Cambria, E., Poria, S., Hazarika, D., Kwok, K. (2016). SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. *Proceedings of COLING 2016*, Osaka, Japan, pp. 2666–2677.
2. Derczynski, L. (2016). Complementarity, F-score, and NLP evaluation. *Proceedings of LREC 2016*, pp. 261–266.
3. Dulău, T. M., Dulău, M. (2019). Cryptocurrency sentiment analysis in social media. *Acta Marisiensis, Seria Technologica*, Vol. 16, pp. 1–6.
4. Golnari, A., Komeili, M. H., Azizi, Z. (2024). Probabilistic deep learning and transfer learning for robust cryptocurrency price prediction. *Expert Systems with Applications*, Vol. 255, pp. 124404.
5. Islam, M. S., Bashir, M., Rahman, S., Al Montaser, M. A., Bortty, J., Nishan, A., Haque, M. R. (2025). Machine learning-based cryptocurrency prediction: Enhancing market forecasting with advanced predictive models. *Journal of Ecohumanism*, Vol. 4, pp. 2498–2519.
6. Kumar, D., Christopher, D., Balamurali, A., et al. (2024). Stock and cryptocurrency price prediction using SVM and LSTM algorithms. *2024 4th Asian Conference on Innovation in Technology (ASIANCON)*, IEEE, pp. 1–5.
7. Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.
8. Otabek, S., Choi, J. (2024). From prediction to profit: A comprehensive review of cryptocurrency trading strategies and price forecasting techniques. *IEEE Access*.
9. Pellicani, A., Pio, G., Ceci, M. (2025). CARROT: Simultaneous prediction of anomalies from groups of correlated cryptocurrency trends. *Expert Systems with Applications*, Vol. 260, pp. 125457.
10. Qureshi, S. M., Saeed, A., Ahmad, F., Khattak, A. R., Almotiri, S. H., Al Ghamdi, M. A., Rukh, M. S. (2025). Evaluating machine learning models for predictive accuracy in cryptocurrency price forecasting. *PeerJ Computer Science*, Vol. 11, pp. e2626.
11. Roelleke, T., Wang, J. (2008). TF-IDF uncovered: A study of theories and probabilities. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 435–442.
12. SenticNet (2024). Sentic API. Accessed 2 July 2024.
13. Shahiki Tash, M., Ahani, Z., Tash, M., Kolesnikova, O., Sidorov, G. (2024). Analyzing emotional trends from X platform using SenticNet: A comparative analysis with

cryptocurrency price. *Cognitive Computation*, Vol. 16, pp. 3168–3185.

14. **Shahiki Tash, M., Kolesnikova, O., Ahani, Z., Sidorov, G. (2024).** Psycholinguistic and emotion analysis of cryptocurrency discourse on X platform. *Scientific Reports*, Vol. 14, pp. 8585.
15. **Sizan, M. M. H., Das, B. C., Shawon, R. E. R., Rana, M. S., Al Montaser, M. A., Chouksey, A., Pant, L. (2023).** AI-enhanced stock market prediction: Evaluating machine

learning models for financial forecasting in the USA. *Journal of Business and Management Studies*, Vol. 5, pp. 152–166.

16. **Vieitez, A., Santos, M., Naranjo, R. (2024).** Machine learning Ethereum cryptocurrency prediction and knowledge-based investment strategies. *Knowledge-Based Systems*, Vol. 299, pp. 112088.

Article received on 20/04/2026; accepted on 23/97/2026.

**Corresponding author is Olga Kolesnikova.*